LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-678126

Kickstarting the Commons: The YFCC100M and the YLI Corpora

J. Bernd, D. Borth, C. Carrano, J. Choi, B. Elizalde, G. Friedland, L. Gottlieb, K. Ni, R. Pearce, D. Poland, K. Ashraf, D. Shamma, B. Thomee

October 13, 2015

ACM Multimedia 2015
Brisbane, Australia
October 26, 2015 through October 30, 2015

Disclaimer

Kickstarting the Commons: The YFCC100M and the YLI Corpora

Julia Bernd[1], Damian Borth[2], Carmen Carrano[3], Jaeyoung Choi[1], Benjamin Elizalde[1], Gerald Friedland[1], Luke Gottlieb[1], Karl Ni[3], Roger Pearce[3], Doug Poland[3], Khalid Ashraf[4], David A. Shamma[5], and Bart Thomee[5]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[3] Lawrence Livermore National Laboratory, Livermore, CA, USA
[4] University of California–Berkeley, CA, USA
[5] Yahoo Labs, San Francisco, CA, USA

[1] {jbernd,jaeyoung,benmael,fractor,luke}@icsi.berkeley.edu
[2] damian.borth@dfki.de
[3] {carrano2,ni4,rpearce,poland1}@llnl.gov
[4] ashrafkhalid@berkeley.edu
[5] {shamma,bthomee}@yahoo-inc.com

ABSTRACT

The publication of the Yahoo Flickr Creative Commons 100 Million dataset (YFCC100M)—to date the largest open-access collection of photos and videos—has provided a unique opportunity to stimulate new research in multimedia analysis and retrieval. To make the YFCC100M even more valuable, we have started working towards supplementing it with a comprehensive set of precomputed features and high-quality ground truth annotations. As part of our efforts, we are releasing the YLI feature corpus, as well as the YLI-GEO and YLI-MED annotation subsets. Under the Multimedia Commons Project (MMCP), we are currently laying the groundwork for a common platform and framework around the YFCC100M that (i) facilitates researchers in contributing additional features and annotations, (ii) supports experimentation on the dataset, and (iii) enables sharing of obtained results. This paper describes the YLI features and annotations released thus far, and sketches our vision for the MMCP.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

MMCommons'15, October 30, 2015, Brisbane, Australia.

ACM 978-1-4503-3744-1/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2814815.2816986.

Keywords

Multimedia; datasets; annotations; YFCC100M; YLI.

1. INTRODUCTION

Building accurate, generalizable, and scalable multimedia analysis and retrieval systems requires large amounts of annotated multimedia data to serve as ground truth. A unique opportunity to meet this need emerged in 2014 with the release of the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset [16], which contains the metadata for approximately 99.2 million photos and 0.8 million videos. To our knowledge, the YFCC100M is the only freely available dataset at this scale, enabling anyone unprecedented access to large amounts of image and video data.

However, to make the dataset even more valuable for multimedia research, it needs to be accompanied by precomputed features and reliable human-generated annotations. To address this need, we are releasing the public-domain YLI feature corpus, a large collection of commonly used visual, aural, and motion features for the YFCC100M dataset. We are also currently developing sets of pilot annotations to stimulate new research in multimedia analysis and retrieval. Two annotated subcorpora, YLI-GEO and YLI-MED, are already publicly available.

Our goal is to use these extensions of the YFCC100M as the seed for a new collaborative endeavor called the Multimedia Commons Project (MMCP). The MMCP aims to provide a common platform and framework for research, supporting and supported by a broad segment of the multimedia community.

This paper first briefly describes the YFCC100M dataset (§2), and then presents the YLI feature corpus (§3), as well as the YLI-GEO (§4) and YLI-MED (§5) annotation subsets, and finally sketches our vision for the Multimedia Commons Project (§6).

2. THE YFCC100M

To meet the need for scale, diversity, and openness in research datasets, Yahoo Labs created the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M), which they released under their Webscope[1] program. YFCC100M is the largest multimedia collection ever released publicly, containing metadata for approximately 99.2 million photos and 0.8 million videos. All were uploaded to Flickr by their creators and published under a Creative Commons[2] (CC) commercial or non-commercial license.

The YFCC100M contains metadata for each of the photos and videos, including user, camera model, timestamp, and machine tags, as well as the title, tags, and text descriptions supplied by the user. About half of the photos and videos also have associated GPS location information. Each item further includes a link to the actual photo or video on Flickr, and to the CC license it was published under. Details about the YFCC100M may be found in Thomee *et al.* 2015 [16]. The original photo and video data are being made available separately via Amazon Web Services (see §3.2).

To make the dataset more accessible to a broader audience, we are developing a YFCC100M browser [6] to provide a real-time retrieval and visualization mechanism for the entire dataset. This online browser allows researchers to explore the dataset by keyword, and augments the retrieved results with statistics such as co-occurrent tags, distribution across users, geo-locations, and time distribution. This analysis can be used for query refinement, and the final index of retrieved images and videos can be downloaded for further processing.

3. THE YLI FEATURE CORPUS

Multimedia analysis relies on numerical descriptors that represent quantifiable features of the multimedia documents, from color histograms and pitch maps to more abstract notions like motion trajectories. Most researchers using a dataset will use similar sets of features extracted using the same tools, creating unnecessary redundancy.

To address this, we are releasing the YLI feature corpus[3], a set of precomputed features extracted from all of the photos and videos in the YFCC100M, as a public-domain resource under the Creative Commons Zero[4] (CC0) license. We are leveraging LLNL's Cray Catalyst high-performance computing infrastructure to provide these features. Our aim is to let the research community focus on advancing the frontiers of science, rather than having them spend time and computational resources on basic tasks like computing features. Over time, the YLI feature corpus will expand in size as we add additional features.

3.1 Included Features

The YLI includes visual, aural, and motion features commonly used in multimedia analysis. The visual features include global (e.g., Gist), local (e.g., SIFT), and texture (e.g., Gabor) descriptors; the aural features include power spectrum (e.g., MFCC) and frequency (e.g., Kaldi) descriptors; and the motion features include dense trajectories [3, 11].

We are extracting a mixture of well-established handmade features and recent, empirically derived deep neural network features, using publicly available tools. For example, the deep-network features originate from CaffeNet [4], a variation on AlexNet [7] computed with Caffe [5].[5]. All of the scripts we used, along with the specific parameters for the YLI features, are available on the YLI website.

Table 1 lays out the full list of features that are currently available for the whole YFCC100M dataset, as well as features we are planning to make available in the future or are considering. Visual features are being computed on both the static images and on keyframes from the videos, with the keyframes captured at 1s intervals. Some features are currently available only for subsets of the data, as described in §4 and §5.

3.2 Distribution Platforms

The computed features are currently being distributed in two ways. Our principal platform[6] provides access to the YLI feature corpus, the YLI-GEO and YLI-MED annotation subsets, and useful tools. The data is hosted by Amazon through their Public Data Sets[7] program on an S3 data store, where we have also made the photo and video data for the YFCC100M available.

Researchers can download the data from AWS for free to their local compute system, or they can access it through subscription services such as Amazon Elastic Compute Cloud (Amazon EC2) or Amazon Elastic MapReduce (Amazon EMR). These services allow researchers to access the data seamlessly without needing to download it to local storage. These services provide highly scalable, on-demand compute that can be used for the analysis of large data collections. Since the YFCC100M metadata is also currently available via the Yahoo Webscope program on their Amazon S3 data store, both can be mounted simultaneously to connect YFCC100M with YLI.

Most of the feature data and preliminary annotations have been mirrored to our secondary platform[8], hosted using Google Drive, from which they can also be freely downloaded. This platform also hosts related multimedia-analysis tools created at ICSI and University of California–Berkeley, including audioCaffe[9], a deep neural net–based audio content–analysis tool [1].

In addition to the features computed by LLNL and ICSI for the public-domain YLI feature corpus, other research groups are also extracting features from the YFCC100M. For example, Popescu *et al.* (2015) have computed VLAD and VGG features for the photos and plan to make them publicly available soon [13]. They are also using these features as the basis for a visual search index[10].

4. YLI-GEO

Subsets of the YLI feature corpus have already been used in the MediaEval Benchmarking Initiative's Placing Task [3]

[1] http://webscope.sandbox.yahoo.com
[2] http://www.creativecommons.org
[3] http://www.yli-corpus.org/computed-features
[4] http://creativecommons.org/publicdomain/zero/1.0/

[5] The differences between CaffeNet and AlexNet are that relighting augmentation is not used and the order of normalization and pooling are reversed.
[6] http://www.mmgenome.org/
[7] https://aws.amazon.com/public-data-sets/
[8] http://www.yli-corpus.org/computed-features
[9] https://github.com/ashrafk/audioCaffeInitial
[10] http://mklab.iti.gr/project/visual-features-and-search-index-flickr-100m-corpus

Data Stream	Feature	Parameters and Notes	Status (8/2015)
Visual (For Photos and for Video Keyframes)	LIRE Features: Auto Color Correlogram, Basic Features, CEDD (Color and Edge Directivity Descriptor), Color Layout, Edge Histogram, FCTH (Fuzzy Color and Texture Histogram), Fuzzy Opponent Histogram, Gabor, Joint Histogram, Joint Opponent Histogram, Scalable Color, Simple Color Histogram (RGB), Tamura Texture	Computed using the LIRE package [10]	Available
	Gist	Computed using Lear's implementation [8]	Available for images; in process for keyframes
	SIFT	Computed using OpenCV [14]	In process
	CaffeNet Features	Layer 6, layer 7, and posterior probabilities; computed using CaffeNet [4], a variant of AlexNet [7]	Prospective
Audio (For Videos)	MFCC20s (Mel Frequency Cepstral Coefficents)	Nineteen lowest channels, plus energy; frame size 25ms; step size 10ms	Available
	Kaldi pitch	Frame size 25ms; step size 10ms	Available
	SAcC pitch (Subband Autocorrelation Classification pitch tracker)	Frame size 25ms; step size 10ms.	Available
Motion (For Videos)	Dense Trajectories	Computed using Lear's implementation with default parameters: trajectory length 15 frames, sampling stride 5 pixels [17]	Prospective

Table 1: Visual, audio, and motion features included in or planned for the YLI corpus.

in 2014 and 2015, and additional YLI data will be added in 2016[11]. These three releases are named MP14, MP15, and MP16, respectively. Each indexes a subset of geotagged YFCC100M metadata, and additionally encompasses computed features for the corresponding photos and videos. The benchmark evolves every year to accommodate the needs of the research community, and thus the composition of the evaluation data and the evaluation criteria change each year as well. We have bundled the three releases into YLI-GEO, so that researchers can easily compare their systems against existing or future results published using MP14, MP15, or MP16 as appropriate.

About half the YFCC100M images and videos have geotags. A subset of these geotagged videos was used to create MP14, by semi-randomly selecting 5 million photos and 25,000 videos for the training set, and 500,000 photos and 10,000 videos for the test set. To ensure variety, we constrained the selection such that no user contributed more than 250 photos and 50 videos; recordings from a given user were each made more than 10 minutes apart; and no user contributed to both the training and test sets. In all, the dataset contains metadata for photos and videos taken by more than 223,000 users. In addition, 1 million photos and 80,000 videos were reserved for the test sets to be provided for the Placing Task in 2015 and 2016.

In 2015, the task shifted focus to the realm of human geographical understanding; the participants' systems were tasked with predicting in which neighborhood, city, state, etc. the photos and videos had been captured. The benchmark also incorporated the domain of human mobility: given a sequence of photos taken in the same city, systems were asked to estimate the locations of those photos that did not have a geotag.

This human-geography dataset, MP15, includes approximately the same training set as MP14, except that some items were filtered out that could not be mapped to an administrative region, for instance those that were captured in international waters/airspace, or some other area not covered by the Database of Global Administrative Areas (GADM)[12]. The training set for MP15 was thus slightly reduced, to 4,695,149 items, of which 4,672,382 are photos and 22,767 videos. The test set for MP15 includes the test set for MP14 plus the additional test set that was set aside for 2015. In total, the MP2015 test set includes 949,889 items, of which 931,573 are photos and 18,316 videos. As before, the training and test sets are disjoint in terms of users.

The availability of the public-domain YLI feature corpus, along with the YFCC100M metadata index, has reduced the data-processing burden on benchmark participants, allowing them to focus on developing innovative approaches to using the feature data for location estimation.

5. YLI-MED

[11] http://www.yli-corpus.org/mediaeval-2014-placing-task-dataset

[12] http://www.gadm.org

YLI-MED is the YLI Multimedia Event Detection corpus[13], a public-domain index of videos with annotations and computed features [2]. The videos in YLI-MED form a subset of the YFCC100M. YLI-MED is specialized for use in multimedia event detection (MED) research, i.e., for training and testing systems that automatically identify particular human-defined events depicted in a video by analyzing its audio and visual content.

5.1 Overview and Rationale

The first edition of YLI-MED contains videos that have been categorized as depicting one of ten target events, or as not showing any of the target events. In addition to annotator-agreement scores and annotator-confidence scores for each categorization, the videos are also annotated with other basic attributes of interest, such as language spoken and production features like musical scoring.

YLI-MED includes ten of the fifteen events covered in the TRECVID MED 2011 [12] evaluation.[14] Our motivations in choosing parallel events were twofold. We wished to provide an alternative dataset to TRECVID MED for which access does not depend on participation in a specific evaluation, and we also wanted to encourage researchers that have already participated in TRECVID MED to test the generalizability of their systems using the target events the two datasets have in common.

We do note that the YLI-MED specifications and collection procedures differed from those for the HAVIC corpus [15] used in TRECVID, in several ways. For example, the event videos in YLI-MED may have non-human protagonists, animated videos were included but videos with no audio track were not, and we did not individually review each of the videos in the negatives set[15].

Bernd *et al.* [2] provide an exhaustive description of the collection procedures, detailed descriptive statistics about the corpus, and a discussion of possible biases in the corpus introduced by our procedural choices, along with a comparison with the HAVIC/TRECVID MED corpus.

5.2 Corpus Composition

The first edition of YLI-MED contains 50,638 videos, including 1,823 videos judged to be "positive instances" of event categories, along with 115 evidential "near misses", and 62 videos "related" to those events but not actually depicting them. There are also 48,638 "negative" videos that do not depict any of the events. Table 2 shows the final numbers of videos for each event category in the first release, divided into standard training and test sets. No single user has videos in both the training and test sets for the same event, and for most events, no single user contributed more than 5% of the videos in either the training or test set.

Each video indexed and annotated by the corpus-collectors was corroborated by at least two additional annotators. The great majority (92.7%) of the "positive" videos for an event were judged as positive instances by all three annotators that viewed them. Average annotator confidence for these videos was 2.71 (on a scale from 1 to 3), while the average confidence for videos where only two out of three annotators agreed was 1.95 (for the positive judgments). However, we observed significant variation in agreement and confidence among the different events. Agreement and confidence scores are included in the public release of the corpus.

Among the positive-example videos in YLI-MED, 1.3% contain animation or CGI, 4.3% have text added in post-production (e.g., titles or subtitles), and 3.9% have music tracks added in post-production. Among the videos with musical scoring, 54% have the score overlaid on top of the original audio, 9% have it interspersed with the original audio, and for 37%, the original audio (if any) was entirely replaced by the musical score.

Spoken (or sung) language is present in 97.0% of the positive videos. Among those videos with spoken/sung language, 90.2% are completely or partially in English, 9.5% are entirely in a non-English language, and 0.4% were unintelligible. Among the videos with text added in post-production, 89.3% have English text.

Among these characteristics, musical scores that replaced audio tracks had a significant negative effect on annotator agreement levels and confidence, while the presence of spoken or sung language had a significant positive effect [2]. (Other production and language characteristics did not seem to affect annotator judgments.)

5.3 Corpus Comparison

To get a preliminary notion of the impact of the (intentional and unintentional) differences between the YLI-MED and TRECVID/HAVIC [15] datasets, we compared results for an i-vector-based audio-analysis system trained and tested on YLI-MED with the same system trained and tested on TRECVID MED data in past work [9]. The system had significantly different performance for many of the events across the two datasets. However, taking all of the events together, the performance differences were not significant (in runs that did not include negative examples) or were only marginally significant (in runs that did include the negatives) [2, 1].

Of course, this is only one specific system, and additional comparison is certainly needed. Still, the performance differences we observed for individual events may indicate that even large corpora like YLI-MED and HAVIC/TRECVID MED may not yet be sufficient for building truly generalizable models.

5.4 Features for YLI-MED

To make the YLI-MED corpus more immediately usable, we are releasing separate bundles of computed features for all videos in this dataset. As of the date of writing, we have publicly released Kaldi pitch features, SAcC pitch, and MFCCs for audio, and CaffeNet/AlexNet and LIRE image features calculated on keyframes, for all of the positive videos in YLI-MED.

6. TOWARD A MULTIMEDIA COMMONS

The efforts currently underway are only the first step. To fully leverage the potential of the YLI feature corpus to transform the state of the art, the multimedia community needs comprehensive annotations for each photo and video in the YFCC100M, spanning computed visual, aural, and motion features and labels for attributes relevant to multiple strands of research.

[13] `http://www.yli-corpus.org/the-yli-med-corpus`

[14] The remaining five events were not sufficiently represented in the YFCC100M.

[15] Around 1% of the negative videos were reviewed by a human annotator to confirm they are not actually positive examples. The other 99% were screened for textual metadata referring to our target events.

Event	Event Name	Train	Test	Total	Avg. Confidence
Ev101	Birthday Party	100	137	237	2.88
Ev102	Flash Mob	100	50	150	2.34
Ev103	Getting a Vehicle Unstuck	100	41	141	2.53
Ev104	Parade	100	130	230	2.83
Ev105	Person Attempting a Board Trick	100	94	194	2.71
Ev106	Person Grooming an Animal	100	39	139	2.76
Ev107	Person Hand-Feeding an Animal	100	120	220	2.87
Ev108	Person Landing a Fish	100	43	143	2.55
Ev109	Wedding Ceremony	100	119	219	2.86
Ev110	Working on a Woodworking Project	100	50	150	2.42
	Total Positives	1000	823	1823	2.71
Ev100	None of the Above	5,000	43,638	48,638	N/A

Table 2: Number of positive-example videos for each event in YLI-MED Version 1.0, and number of negative videos, with average confidence scores for all positive judgments.

We are therefore laying the groundwork for a common platform and framework around the YFCC100M, which we call the Multimedia Commons Project (MMCP). With the MMCP, we aim to bring together researchers in image and video analysis around creating publicly available features and annotations for the YFCC100M—and thereby to stimulate the formation of new research directions and collaborations.[16]

6.1 The Value of Common Datasets

While over time the techniques used in multimedia analysis may change, researchers generally will need reliable annotations—and especially so for consumer-produced media, which can vary wildly in subject matter, style, and quality. Large, open-access annotated corpora, such as YLI, intend to enable comprehensive, generalizable approaches.

For example, one strand of the Multimedia Commons focuses on the development of a "multimedia genome" in the form of a reliable set of annotations for all of the videos in the YFCC100M. We anticipate this will foster a fundamental understanding of the underlying structure of consumer-produced videos, enabling researchers to uncover the major elements contained in consumer-produced videos, identify patterns in their typical structures, and determine what visual and acoustic characteristics make them searchable.

Openness is key to advancement, by encouraging broad participation and allowing researchers to evaluate and reproduce results obtained by others. There are understandable reasons why the largest and most reliably annotated datasets tend to not be available to the public without restrictions or fees, such as the effort of collecting and annotating large amounts of data. However, barriers to access impede research progress, especially for students and for researchers at small institutions that do not have the resources to purchase corpus subscriptions or participate in evaluations. Having freely accessible and high-quality common datasets, on the other hand, helps to level the playing field.

[16]A note on names: As other research groups begin to contribute features and annotations, this collective resource (including the YLI and other groups' data) will be named by the community. For the present, we are calling the prospective data collection the "Multimedia Commons Dataset", and the set of projects and initiatives to gather and analyze that data the "Multimedia Commons Project". However, this is not fully settled; for example, the names "Multimedia Genome Dataset" and "Multimedia Genome Project" have also been suggested as cover terms.

6.2 Community-Driven Annotation

We believe it is important for a project that aims to support and be supported by a research community to closely involve as many participants as possible. To this end, we have been, and still are, soliciting input from multimedia researchers—including computer vision, image analysis, audio, multimedia, and speech researchers—about their priorities for a large, open-source, richly annotated photo and video corpus. We are particularly interested in engaging researchers whose ability to perform research on YFCC100M is resource-contrained (e.g., local computation of features is not feaasible), so that we can explore how best to mitigate those constraints.

Some inputs that are required in order to shape near-term decisions for MMCP include how to optimize depth vs. coverage in annotations, which formats the data should be in, and how to define annotation parameters and label vocabularies to make the data most useful for a broad set of tasks. For example, annotation decisions might be made along dimensions such as complexity (e.g., low-level concepts/percepts vs. high-level events), spatial resolution (e.g., whole images vs. bounding boxes), objectivity vs. subjectivity (e.g., objects vs. affect), or temporal resolution (e.g., scene vs. shot).

6.3 Enabling New Research Directions

The YFCC100M and the YLI feature corpus have already been used in several benchmarks and challenges. But with a coordinated community effort to construct, over time, new annotations and challenge problems, they have the potential to enable new directions in multimedia research.

In particular, there is much untrodden ground in the trade spaces enumerated in the previous section, and it is our belief that the creation of recognized metrics in new regions of those trade spaces (e.g., annotation of video events at multiple spatial, temporal, and abstraction levels) will spur new research and new developments.

In addition to inspiring new, independent benchmarks and challenges, the MMCP could be the catalyst for the creation of a larger platform for multimedia researchers to run experiments and share results. The MMCP will have the most impact if multiple research groups around the world con-

tribute annotations and features they are creating for their research, thus providing a diversity of ground-truth datasets. Collaborative research will in turn keep the corpus dynamic, as annotations expand to cover new areas of interest for the field.

7. ACKNOWLEDGMENTS

Work on the YLI corpus and the Multimedia Commons Project is supported by several funders, including: a collaborative Laboratory Directed Research and Development project led by Lawrence Livermore National Laboratory, under the auspices of the U.S. Dept. of Energy contract DE-AC52-07NA27344 (LLNL-CONF-676635); a grant from Cisco Systems, Inc. for Event Detection for Improved Speaker Diarization and Meeting Analysis; and a National Science Foundation grant for the SMASH project: Scalable Multimedia content AnalysiS in a High-level language (award IIS-1251276). (Any opinions, findings, and conclusions expressed here are those of the individual researchers, and do not necessarily reflect the views of the funders.)

We would also like the thank the following contributors to the various YLI-related efforts described in this paper: *YFCC100M:* Li-Jia Li and Nikhil Rasiwasia; Yahoo Webscope and Legal teams. *The YFCC100M Browser:* Andreas Dengel, Sebastian Kalkowski, and Christian Schulze. *YLI-MED:* Heather Gallagher, Adam Janin, Sara Karabashlieva, Florin Langer, Jocelyn Takahashi, and Jennifer Won. *YLI @ AWS Public Data Sets:* Ariel Gold, Matt Jamieson, and KD Singh. *YLI Distribution:* Karina Goot and Adam Janin. *Multimedia COMMONS Workshop:* Martha Larson and Chong-Wah Ngo.

8. REFERENCES

[1] K. Ashraf, B. Elizalde, F. Iandola, M. Moskewicz, G. Friedland, K. Keutzer, and J. Bernd. Audio-based multimedia event detection with DNNs and sparse sampling. In *Proceedings of the 5th ACM International Conference on Multimedia Retrieval (ICMR '15)*, 2015.

[2] J. Bernd, D. Borth, B. Elizalde, G. Friedland, H. Gallagher, L. Gottlieb, A. Janin, S. Karabashlieva, J. Takahashi, and J. Won. The YLI-MED corpus: Characteristics, procedures, and plans (ICSI Technical Report TR-15-001). *arXiv:1503.04250*, 2015.

[3] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, and D. Poland. The Placing Task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of the ACM Multimedia 2014 Workshop on Geotagging and Its Applications in Multimedia (GeoMM '14)*, Orlando, FL, November 2014. Association for Computing Machinery.

[4] J. Donahue. Caffenet model from modelzoo. `https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet`, 2012.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[6] S. Kalkowski, D. Borth, C. Schulze, and A. Dengel. Real-time analysis and visualization of the YFCC100M dataset. In *Proceedings of the ACM Multimedia 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions (MMCommons '15)*, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[8] LEAR. Lear's GIST implementation. `http://lear.inrialpes.fr/software`.

[9] J. Liu, H. Cheng, O. Javed, Q. Yu, I. Chakraborty, W. Zhang, A. Divakaran, H. S. Sawhney, J. Allan, R. Manmatha, J. Foley, M. Shah, A. Dehghan, M. Witbrock, J. Curtis, and G. Friedland. SRI-Sarnoff AURORA system at TRECVID 2013: Multimedia event detection and recounting. In *TREC Video Retrieval Evaluation: Notebook Papers and Slides*, 2013.

[10] M. Lux and O. Marques. *Visual Information Retrieval using Java and LIRE*. Morgan & Claypool, San Rafael, CA, 2013.

[11] K. S. Ni, C. C. Carrano, D. N. Poland, B. M. Elizalde, G. Friedland, L. R. Gottlieb, and D. S. Borth. The Yahoo-Livermore-ICSI (YLI) multimedia feature set. Technical Report LLNL-MI-659231, Lawrence Livermore National Laboratories, August 2014.

[12] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. QuÃl'not. TRECVID 2011 - an overview of the goals, tasks, data, evaluation mechanisms, and metrics. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, May 2012.

[13] A. Popescu, E. Spyromitros-Xioufis, S. Papadopoulos, H. L. Borgne, and Y. Kompatsiaris. Toward an automatic evaluation of retrieval performance with large scale image collections. In *Proceedings of the ACM Multimedia 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions (MMCommons '15)*, 2015.

[14] K. Pulli, A. Baksheev, K. Kornyakov, and V. Eruhimov. Real-time computer vision with OpenCV. *Communications of the ACM*, 55(6):61–69, 2012.

[15] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating HAVIC: Heterogeneous audio visual Internet collection. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[16] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2015. To appear.

[17] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV 2013 - IEEE International Conference on Computer Vision*, pages 3551–3558, Sydney, Australia, Dec. 2013. IEEE.